4-1-2013

Waiting for Weighted Navigation

David Stern
Illinois State University, hstern2@ilstu.edu

Recommended Citation
"Waiting for Weighted Navigation" Online Searcher 37 (2): 51-55 (March/April 2013).

Waiting for Weighted Navigation

by David Stern

How effective are our aggregated discovery search tools? Are the navigation, filtering, and effectiveness of preliminary federated and/or harvesting search tools adequate to overcome the swamping effect that can occur when quantity obscures quality or entire classes of information are overlooked? One example, noted by Jevin D. West in his Ph.D. dissertation ("Eigenfactor: Ranking and Mapping Scientific Knowledge"; octavia.zoology.washington.edu/people/jevin/Documents/Dissertation_JevinWest.pdf), is of larger journals swamping smaller ones. Another example of the swamping effect happens when searching large domain subject indexes, compared to searching MARC records within the OPAC. The former can easily swamp the latter.

This article will provide a study of some technical navigational assistance hedges that will lead to enhancements in existing and future discovery and use. Currently, it seems that the techniques provided in these aggregator search tools are not always adequate to assist uninitiated users in discovering many of the quality materials that lie deep within the aggregated content. This is demonstrated by the minimal discovery of quality OSTI (Office of Scientific & Technical Information, U.S. Department of Energy) materials by key research populations when compared to surprisingly and significantly greater use of these same high-quality primary technical data by lower-level research institutions.

The OSTI data use implies that emphasized records within a smaller domain of OPAC materials lead to more effective/appropriate access to quality material than access to those same high-quality materials discovered through the use of sophisticated aggregated content search tools. My research provides startling data about the high rate of failure to identify significant Department of Energy (DOE) materials within research environments when the materials are within the searchable domain.

What is missing are methods of displaying results that offer users guidance in clustering results based upon common weighted variables. The new aggregated search technologies would be improved by emulating the behaviors learned by experienced researchers and enhancing these refinement and suggestion methods with new analytical techniques.

HOW EXPERIENCED SEARCHERS BEHAVE

The experienced searcher employs a set of discipline-specific methods developed over time and shared via colleagues to narrow search result sets. These selection

techniques might include refining results based upon 1) known quality journals, 2) known key authors, 3) leading institutions, 4) publications and presentations from important conferences, and 5) specific government agencies of known relevance.

Some advanced researchers go beyond these personal knowledge steps, using added-value navigation tools such as citation tracking and Find Similar algorithms. These important next-level navigational tools use computer power to find additional articles. However, they generally add new results; they do not target or refine result sets using qualitative assessment. Even some of the more controversial navigation options offered, such as bX Recommender (which provides usage-based suggestions from aggregated SFX links), add new materials to consider rather than evaluating the existing result set.

COMPUTER-BASED EVALUATION METHODS

Computer processing can now assist in this important evaluation and selection task, providing assistance with quality results identification. Some of these metrics include the Eigenfactor, Article Influence Score, Thomson Reuters JCR, SJR (SCImago Journal and Country Rank), and SNIP (Source Normalized Impact per Paper).

In the Eigenfactor (eigenfactor.org) approach, a citation in a high-profile journal, such as *Nature*, counts for more than a citation from a journal that few people read or cite. It eliminates self-citations and weighs each reference according to a measure of the time researchers spend reading the journal itself. To many people, this method provides a better way to evaluate a journal's quality than the widely used impact factor, which tracks how many citations a journal title receives, but it does not weight the quality of the related sources.

A significant example of a viable method for developing a weighted tool was recently announced in "Citation by Citation, New Maps Chart Hot Research and Scholarship's Hidden Terrain," *Chronicle of Higher Education*, Sept. 11, 2011, by Jennifer Howard (chronicle.com/article/Maps-of-Citations-Uncover-New/128938/?sid=at&utm_source=at&utm_medium=en).

She describes how a team of researchers led by two biologists, Carl T. Bergstrom and Jevin D. West, and a physicist, Martin Rosvall, set out to build just such a guidance system. They intend to use the Eigenvalue measurement, a now-common evaluation tool based upon the relative quality of the source materials, to create new ways to identify and analyze patterns in millions of journal citations. The key element is the use of a quality measurement, adding refinements by calculating the quality of the source and citation sources themselves.

Regardless of the perfection of the assessment tool, which is still open to debate, there is now the ability to provide some refinement assistance as a major step toward weighted navigation.

The Article Influence Score is also from the Bergstrom team; it uses Thomson Reuters JCR citation data to calculate the relative importance of a journal on a per-article basis. It is the Eigenfactor score divided by the fraction of articles published by the journal, providing a score normalized above or below 1.0 for influence.

SJR and SNIP are two other examples of journal evaluation methods based upon Elsevier citation data factored across the total number of citations for the journal during the previous 3 years. While not as complex and qualitative as the Eigenfactor approach, both do provide a way to rank the impact of specific journals and could lead to recommendations.

NEXT STEPS IN ADDING CRITERIA

How effective would a discipline-specific regression analysis routine be if we could provide weighted values to the earlier mentioned personal characteristics used by experienced searchers? For this study in particular, the weighting of quality government document resources from specific agencies might raise the visibility of the hidden materials that are not being found. Given that government documents are frequently not part of a library's OPAC, you could consider them the "lost children" of the collection.

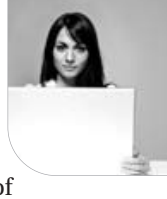
Imagine profiles of weighted variables for subject areas, which could be selected as a limit option by users. These profile criteria could be created by a panel of experts in a field and provided as part of the general interface. There could even be profiles for different levels of users: technical information, theoretical considerations, graduate student overviews, and undergraduate orientation perspectives. Customized slices of appropriate materials could be developed across the domains based upon user expectations and desires.

THE OSTI DATA

OSTI used Google Analytics to review the access statistics for its materials from major libraries. The patterns are troubling for those hoping to see correlations between top science institutions and top database usage. The data shows that those libraries loading MARC records for DOE materials have *significantly* (as in orders of magnitude) greater retrievals than the largest research libraries, which are assumed to have the appropriate subject indexes, with most offered via federated search tools.

Searchers starting directly in science.gov or the Information Bridge interface (osti.gov/bridge), which provides free public access to almost 300,000 full-text documents and bibliographic citations of DOE research report literature, also show far higher use of OSTI materials than those starting in federated or aggregated search tools, once again showing a clear swamping effect. In these cases, it appears that sophisticated topic searching and browsing are done more effectively in smaller domain databases than in the larger aggregated and federated databases, which probably leads to even more comprehensive access than OPAC MARC searching because of the additional access points and facets.

Our data indicates that the federated searches, or single database searches, are not highlighting the DOE records that should be important and discovered by serious researchers



in these premiere institutions. The DOE records are being swamped. Table 1 shows the top 30 institutions by number of page retrievals—eight of the top 10 are institutions with MARC records loaded, and 13 of the top 20 were from institutions with MARC record loads. It would be hard to defend the ranking of top institutions on this list compared to the research intensity of the schools in the DOE areas. One would expect schools 12, 13, and 14 (Stanford University, Massachusetts Institute of Technology, and The Pennsylvania State University) to be at the top of this list based upon need for timely government information, and yet they are eclipsed by other schools due to the apparent MARC/OPAC and/or Information Bridge discovery advantage.

Table 2 contains the ranked information for the top users, and the data demonstrates that those researchers accessing the material from MARC records spend less time accessing and browsing OSTI records, probably due to greater initial precision. The schools without MARC records show much greater browsing of material using the OSTI Information Bridge interface. This may indicate initially less precise or comprehensive search results in the indexing databases and federated searches.

It appears that database searching and federated searching (even within specific subject clusters) do not provide obvious access to all the expected quality government materials due to swamping effects found in large and heterogeneous domain databases through one-stop searching.

The data shows that smaller (less research-intensive) schools with OPAC record loads are far higher users of this important data than the research schools without such record loads. Given the limitations of the large indexing databases and aggregator search tools to recommend the most relevant materials, those intensive research institutions that perform the OPAC load will see far higher use from the OPAC access than from well-designed subject database searching, LibGuide-type database pointers, or federated searching.

THE IMPLICATIONS

Given concerns about the effectiveness of aggregated discovery, can some of the factors be identified?

The first issue is "complexity masks simplicity." Our aggregated interfaces perform many sophisticated processes, but they do not easily offer advanced refinement assistance, even if the possibilities exist within the system. The early emphasis was on performing broad searches that cross disciplines. This service has not yet been well-developed, because sophisticated ontology and vocabulary normalization has not been implemented. Perhaps it is now time to focus some time and attention on the development of equally important post-search refinement assistance tools.

A second concern in using any search system is providing flexibility when dealing with people of multiple intelligences, the tendencies for people to think most comfortably in various ways—textual, visual, auditory, and tactile. Providing powerful tools will probably eventually require various interfaces and search methods to address these preferences and tendencies. For now, most aggregated search interfaces offer few variations in terms of visualizing and refining search result sets. We may have removed the need for Boolean operators, but our return to simple term entry does not maximize natural-language processing and intuitive visualization possibilities. We should concentrate on navigation, not content.

QUESTIONS TO CONSIDER

While we wait for weighted navigation, for enhanced assistance within our aggregator search tools, let's consider some practical questions for libraries and other portals.

1. In the short term, do we train users on aggregated (federated or harvested) tools, or do we add the most common items to the OPAC for immediate discovery? This is particularly pertinent for government information, which users already mistakenly believe are in the OPAC due to historical

Table 1. Top Institutions by Page Retrievals

Information Bridge Page Retrievals by University Libraries, January–October 2011

School	Number of Page Retrievals	Page Retrievals not From Library Catalogs	DOE Office of Science Funding FY 2009 & 2010
University of Florida	8289	28	$13,286,000
Purdue University	5778	194	$13,724,000
University of Missouri	5436	230	$6,503,000
University of South Florida	4020	7	$1,839,000
University of Colorado	2810	85	$20,845,000
Florida International University	2532	264	$0
University of Iowa	2197	25	$3,433,000
Michigan State University	1365	115	$41,524,000
Stanford University	1295		$23,963,000
MIT	1157		$150,420,00
Penn State University	1087		$41,932,000
Georgia Tech	242	808	$10,036,000
University of West Florida	985	0	$0
Auburn University	915	23	$3,432,000
Oklahoma State University	903	32	$2,712,000
University of Massachusetts	818		$34,628,000
University of Texas	794		$42,663,000

Table 2. Top Institutions With Time and Page Browsing

Source	Visits	Pages Viewed	Pages/Visit	Average Time on Site
uf.catalog.fcla.edu	2257	6889	3.052282	276.8977
catalog.lib.purdue.edu	1502	5689	3.787617	222.5333
laurel.lso.missouri.edu	1340	5101	3.806716	257.8754
usf.catalog.fcla.edu	1086	3084	2.839779	232.6013
encore.colorado.edu	621	2226	3.584541	197.5588
smartsearch.uiowa.edu	803	2194	2.732254	169.0523
fiu.catalog.fcla.edu	733	2064	2.815825	324.6726
uf.catalog.fcla.edu.lp.hscl.ufl.edu	475	1400	2.947368	263.56
catalog.lib.msu.edu	442	1292	2.923077	239.5158
usf.catalog.fcla.edu.ezproxy.lib.usf.edu	338	934	2.763314	237.3491
uwf.catalog.fcla.edu	269	911	3.386617	266.2379
libguides.mit.edu	119	882	7.411765	381.0336
library.umass.edu	70	720	10.28571	281.9429
uwcatalog.uwyo.edu	247	692	2.801619	305.9838
lib.utexas.edu	74	686	9.27027	437.7162
catalog.lib.auburn.edu	209	654	3.129187	382.3876
highwire.stanford.edu	92	624	6.782609	243.4457
gtsearch.library.gatech.edu	118	608	5.152542	200.5763
libraries.psu.edu	95	600	6.315789	252.5895
library.lib.asu.edu	67	584	8.716418	726.2388
libraries.colorado.edu	137	569	4.153285	213.5985
osucatalog.library.okstate.edu	156	537	3.442308	223.3397
muse.lib.ncku.edu.tw:8080	52	513	9.865385	335.2308
ju.edu.et	114	488	4.280702	155.7895
fiu.catalog.fcla.edu.ezproxy.fiu.edu	160	468	2.925	205.3875
furbo.gmu.edu	99	387	3.909091	141.1515
an5qy7ag4q.cs.serialssolutions.com .libproxy.cc.stonybrook.edu	73	374	5.123288	303.726
lib.washington.edu	36	370	10.27778	510.1667
library.temple.edu	52	367	7.057692	379.1346
lib.rpi.edu	38	364	9.578947	433.4474
library.lehigh.edu	41	334	8.146341	285.4146
uri.edu	46	328	7.130435	436.8261
library.mines.edu	44	327	7.431818	318.0909
subjectguides.esc.edu	33	319	9.666667	426.303
mobius.missouri.edu	101	306	3.029703	208.1683
mobius.umsystem.edu	92	301	3.271739	137.5217
math.columbia.edu	168	296	1.761905	145.3274
onlinebooks.library.upenn.edu	88	296	3.363636	169.7614
energy.wsu.edu	126	294	2.333333	130.1587
libweb1.lib.buffalo.edu	38	293	7.710526	278.4474
fusion.erau.edu	42	280	6.666667	251.5952

selective Marcive loading. Of course, this assumes we are working with known quality materials with MARC-compatible records.
2. Do we emphasize deep personal exploration of all resources or provide assisted filtering? We find ourselves in the classic conundrum—offering some sort of guided access versus requiring comprehensive and contextual information analysis by the end user. The best place along this continuum probably depends upon where you sit in the information life cycle and how intuitive the interface is for helping users filter to the most appropriate items.
3. Should we wait for the best tool, or should we act now and become involved in developing the ultimate solutions? Should libraries load more data into the OPAC if it works, offer powerful but confusing search tools that require greater understanding for effective use, or hold out for better web discovery methods that are not yet mature?

I believe that customized pages, customized precreated search boxes, and personalized tools will ultimately provide the best comprehensive searching, this does not actually seem to work yet according to the data reviewed by OSTI.

BEYOND ENHANCED ASSISTANCE WITHIN INDEXING SEARCH TOOLS

Another important consideration hovering on the outside of this indexing scenario involves full-text searching. How much longer will index searching, whether it's easy and immediate access to MARC records in the OPAC or enhanced navigation of harvested metadata, remain a better approach than searching more complete records and/or full text?

The same concerns about filtering and precision will exist when dealing with full text, and there will probably need to be a combination of initial full-text searching with limits using metadata facets. The same questions of weighted and customized filtering will remain, and perhaps they will become even more important, as the initial full-text material will have even greater degrees of interpretations and false drops.

Is the technology behind artificial intelligence and normalized ontologies simply not mature enough to handle indexing, never mind the larger scale of full-text materials? Can we expect technological solutions ever to understand and interpret our meanings when creating search algorithms across complex and deep materials? Will we always require a significant amount of personal critical thinking and expansive research exploration to perform effective searches? At what level of depth, and with what level of safety, can we trust computer analysis to provide reasonable assistance?

Will new visual interfaces provide better navigation through these complex domains? Is there something about visual clues that might help (at least a portion of the population) to more easily navigate among the very deep layers of available materials? Can human brains make connections visually that are not possible using textual clues?

How successful will these visual interfaces be in providing assistance with deep searching into media materials? This is a growing domain, and searching and displaying multimedia will require new approaches, new metadata considerations, and even better navigational hedges.

Finally, how long will it take to effectively mine other types of materials such as geographic information systems, datasets, and dynamic databases?

DISCOVERY SERVICES VERSUS FEDERATED SEARCH EFFECTIVENESS

Already full-text discovery search tools such as Serials Solutions Summon, EBSCO Discovery Service, Ex Libris Primo, and WorldCat Local allow people to mine deeper into raw materials than ever before. No longer must we rely on broad metadata descriptors for finding unsearchable content. Searching can now retrieve material previously buried within the text of books and other harvested material.

Soon discovery search systems will be able to refine these search results using more complex and heterogeneous materials in innovative and assistive ways. As these discovery services become more widespread, they should provide even greater access points through full-text mining. It will be interesting to see if their sophisticated searching algorithms can provide better results based on the greater amount of data available to be manipulated. While it is too soon to tell if these tools are successful, it should be possible to observe a migration of OSTI hits away from the direct Info Bridge platform and to the discovery services platforms. We might also see a significant increase in the number of hits. If the OSTI Information Bridge materials are included in the harvested data, the swamping effect is mitigated by better search and refinement tools.

At this time, with only a small amount of time and data available for analysis, only Penn State appears to have both discovery service access and relatively high usage. Perhaps this correlational relationship will be possible to analyze in terms of causation. I look forward to a follow-up analysis to determine if discovery services are providing better recall and precision than aggregated and federated systems. This future analysis will provide some evidence from a parallel test of the power of full-text searching and harvesting possibilities versus index searching.

Many questions remain unanswered, and perhaps even unasked, about providing effective search assistance and navigational tools. Continuing explorations of various indexing, searching, analyzing, and displaying options will lead us to new retrieval approaches and new search facilitation possibilities.

David Stern *(david.stern@ilstu.edu) is associate dean for public services, Illinois State University.*

Comments? Email the editor-in-chief (marydee@xmission.com).